Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Fully Automated Segmentation of Head CT Neuroanatomy Using Deep Learning

Jason C. Cai, MD • Zeynettin Akkus, PhD • Kenneth A. Philbrick, PhD • Arunnit Boonrod, MD • Safa Hoodeshenas, MD • Alexander D. Weston, PhD • Pouria Rouzrokh, MD, MPH • Gian Marco Conte, MD, PhD • Atefeh Zeinoddini, MD • David C. Vogelsang, BS • Qiao Huang, PhD • Bradley J. Erickson, MD, PhD

Departments of Radiology (J.C.C., K.A.P., S.H., P.R., G.M.C., D.C.V., Q.H., B.J.E.) and Cardiovascular Science (Z.A.), Mayo Clinic Rochester, 200 First St. SW, RO_PB_02_RIL, Rochester, MN 55905; Department of Radiology, Khon Kaen University, Khon Kaen, Thailand (A.B.); Department of Health Sciences Research, Mayo Clinic Florida, Jacksonville, Fla (A.D.W.); and Department of Internal Medicine, Ascension St. John Hospital, Detroit, Mich (A.Z.). Received October 28, 2019; revision requested December 17; revision received June 2, 2020; accepted June 16. **Address correspondence to** J.C.C. (e-mail: *jason.ccai@gmail.com*).

Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(5):e190183 • https://doi.org/10.1148/ryai.2020190183 • Content codes: AI CT NR

Purpose: To develop a deep learning model that segments intracranial structures on head CT scans.

Materials and Methods: In this retrospective study, a primary dataset containing 62 normal noncontrast head CT scans from 62 patients (mean age, 73 years; age range, 27–95 years) acquired between August and December 2018 was used for model development. Eleven intracranial structures were manually annotated on the axial oblique series. The dataset was split into 40 scans for training, 10 for validation, and 12 for testing. After initial training, eight model configurations were evaluated on the validation dataset and the highest performing model was evaluated on the test dataset. Interobserver variability was reported using multirater consensus labels obtained from the test dataset. To ensure that the model learned generalizable features, it was further evaluated on two secondary datasets containing 12 volumes with idiopathic normal pressure hydrocephalus (iNPH) and 30 normal volumes from a publicly available source. Statistical significance was determined using categorical linear regression with $P < .05$.

Results: Overall Dice coefficient on the primary test dataset was 0.84 ± 0.05 (standard deviation). Performance ranged from 0.96 ± 0.01 (brainstem and cerebrum) to 0.74 ± 0.06 (internal capsule). Dice coefficients were comparable to expert annotations and exceeded those of existing segmentation methods. The model remained robust on external CT scans and scans demonstrating ventricular enlargement. The use of within-network normalization and class weighting facilitated learning of underrepresented classes.

Conclusion: Automated segmentation of CT neuroanatomy is feasible with a high degree of accuracy. The model generalized to external CT scans as well as scans demonstrating iNPH.

Supplemental material is available for this article.

CT is a cornerstone of neuroimaging, and its use has increased steadily (1). Given the large volume of examinations, fully automated algorithms can potentially augment clinical workflow and improve diagnostic accuracy. Convolutional neural networks (CNNs) are the current standard in medical image analysis, and they can automate routine classification and segmentation tasks (2).

CNNs can accurately perform segmentation at US (3) and MRI (4). Several groups have also trained CNNs to delineate hemorrhage (5,6), tumors (7), whole brain (8), and cerebrospinal fluid (9) from head CT. Although algorithms that segment intracranial anatomy at MRI have been shown to facilitate structural analysis and disease characterization (4,10), there is currently no equivalent method for CT despite its higher utilization (1). Such an algorithm could be useful for a variety of reasons. First, knowledge of anatomy could enable downstream tools to localize pathologic features and predict neurologic outcome. For example, Takahashi et al (11) combined CNN and atlas-based approaches to identify acute ischemic stroke from noncontrast CT. Second, accurate segmentations can guide clinicians in treatment planning, such as radiation therapy and image-guided intervention using CT (12,13). Last, routine large-scale analysis of volumetric and morphologic information may lead to the discovery of new imaging biomarkers. Several studies have already demonstrated the clinical utility of volumetric information in head CT. For example, Diprose et al (14) found that cerebrospinal fluid volumes correlate with reduced functional independence after endovascular thrombectomy for ischemic stroke, and Anderson et al (15) found that lower ventricular volume is associated with clinical improvement following shunt insertion for idiopathic normal pressure hydrocephalus (iNPH). To improve the diagnosis of iNPH, Toma et al (16) further proposed the use of population-based image analytics to establish normal limits for intracranial compartment volumes. However, the adoption of these methods, as well as the development and assessment of new CT imaging features, have been hindered by the lack of an accurate and fully automated segmentation algorithm.

In this study, we trained and validated a U-Net (17) to simultaneously delineate multiple intracranial structures from noncontrast CT. As these structures ranged considerably in size and slicewise prevalence, we described strategies

Abbreviations

BC = brainstem and cerebrum, CA = internal capsule, CE = cerebellum, CN = caudate nucleus, CNN = convolutional neural network, CO = insular cortex, CS = central sulcus, DICOM = Digital Imaging and Communications in Medicine, DV = dural folds and venous sinuses, iNPH = idiopathic normal pressure hydrocephalus, LN = lentiform nucleus, ROI = region of interest, RSNA = Radiological Society of North America, SP = septum pellucidum, SS = subarachnoid space, VS = ventricular system

Summary

Deep learning can accurately segment neuroanatomy on CT scans; optimizations that improved segmentation performance for structures with varying size and slicewise prevalence were assessed, including the use of within-network normalization techniques and class weighting schemes.

Key Points

- A fully automated deep learning model that simultaneously segments 11 intracranial structures at head CT was presented.
- The effects of within-network normalization techniques, loss functions, and class weighting schemes on network training and performance were demonstrated.
- The model generalized to external CT scans and scans demonstrating idiopathic normal pressure hydrocephalus; overall Dice coefficients were comparable to expert annotations and higher than those of existing segmentation methods.

to mitigate challenges associated with multiclass segmentation and class imbalance. To ensure that our model learned generalizable features, we validated it on examinations demonstrating iNPH as well as on normal examinations from the Radiological Society of North America (RSNA) Intracranial Hemorrhage Detection Challenge. Finally, we compared its results with existing segmentation methods.

Materials and Methods

Data Acquisition

Primary dataset.—For this retrospective study, we obtained institutional review board approval with a waiver of informed consent and Health Insurance Portability and Accountability Act authorization. We retrieved a consecutive list of all noncontrast head CT scans performed between August and December 2018, and we excluded pediatric patients and patients with identifiable intracranial abnormalities (except brain parenchymal atrophy and leukoaraiosis). From this list, we randomly selected 62 unique examinations (mean patient age, 73 years; age range, 27–95 years; 35 women) and extracted their axial oblique series. There is no overlap of participants with previous studies.

Secondary test datasets.—We acquired two additional datasets to assess model generalizability. The first dataset contained 30 volumes from the RSNA Intracranial Hemorrhage Detection Challenge. All patients demonstrated varying degrees of brain parenchymal atrophy with no other intracranial abnormalities (no hemorrhage or chronic infarction). The second dataset contained 12 patients (mean age, 74 years; age range, 60–84 years; seven women) evaluated at our center for iNPH between August and December 2018. We intentionally selected only examinations that demonstrated severe ventriculomegaly. Six of these patients also had newly inserted ventricular shunts. We chose iNPH because ventricular enlargement distorts but does not efface normal anatomy. The presence of highly attenuating foreign objects further tested the model's robustness. Table E1 (supplement) summarizes Digital Imaging and Communications in Medicine (DICOM) information for the primary and iNPH datasets. DICOM information was not published for the RSNA dataset.

Data Preparation

Primary dataset.—Patient volumes were split into 40 scans (1738 slices) for training, 10 (397 slices) for validation, and 12 (530 slices) for testing. Scans were segmented into 11 regions of interest (ROIs) by a team of radiography students trained by two radiologists (A.B., a neuroradiologist with 9 years of experience, and S.H., a general radiologist with 7 years of experience). They were subsequently corrected by a medical doctor (J.C.C.) and verified by A.B. Specifically, the 11 ROIs were as follows: brainstem and cerebrum (BC, as one class), caudate nucleus (CN), central sulcus (CS), cerebellum (CE), dural folds and venous sinuses (DV, as one class), insular cortex (CO), internal capsule (CA), lentiform nucleus (LN), septum pellucidum (SP), subarachnoid space (SS), and ventricular system (VS). The supratentorial structures were chosen because they represent regions typically supplied by the middle cerebral artery, the largest cerebral artery and the artery most commonly affected in ischemic stroke (18). Several structures are also components of the Alberta Stroke Programme Early CT Score (19). The transverse sinus was not segmented as it did not appear consistently in axial sections; it was included with either BC or CE. We used RIL-Contour to annotate the dataset (20).

Ground truth masks.—From the primary test dataset, we created a series of multiobserver consensus labels, which we refer to as ground truth masks. Specifically, from each volume we selected four to six slices to represent all 11 ROIs (59 total slices). One slice contained CE, one slice contained CS at its most visible level, and two to four contiguous slices contained CA, CN, CO, LN, and SP. The remaining classes were represented across multiple slices. All slices were independently annotated by A.B., S.H., and J.C.C. From each set of three observations, a ground truth mask was generated by majority voting. If all three observers disagreed, a voxel was deemed indeterminate and was excluded from performance metrics. Indeterminate voxels made up 1.1% of all annotated voxels. Our workflow is outlined in Figure 1.

Secondary test datasets.—The iNPH dataset (12 volumes with 423 slices) was segmented in the same way as the primary dataset, while the RSNA dataset (30 volumes with 150 slices) was segmented in the same way as the ground truth masks,

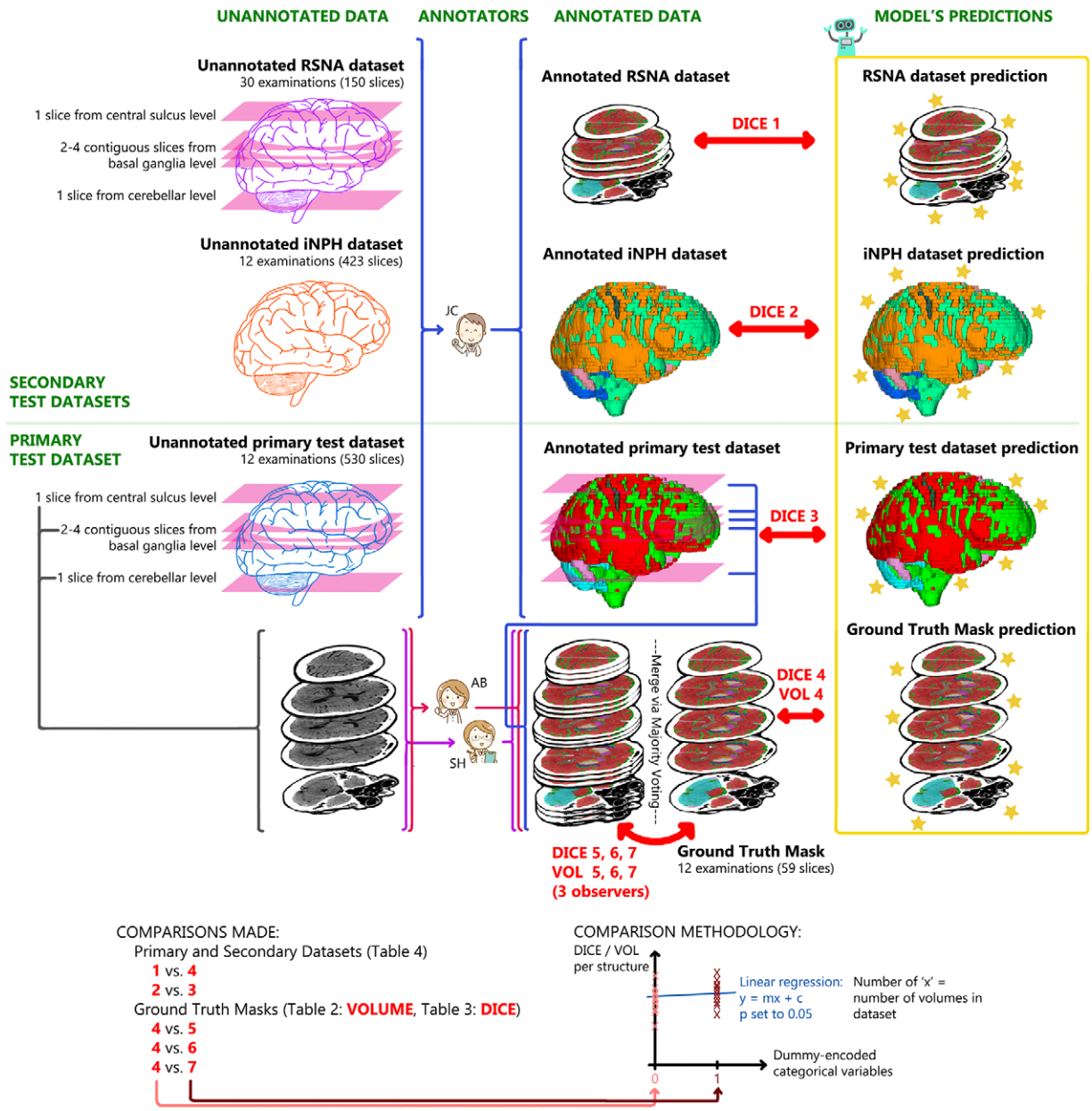

Figure 1: Model evaluation and test dataset workflow. AB = Arunnit Boonrod, MD; DICE = Dice coefficients; iNPH = idiopathic normal pressure hydrocephalus; JC = Jason Cai, MD; RSNA = Radiological Society of North America; SH = Safa Hoodeshenas, MD; VOL = differences in structure volume.

except that images were annotated by J.C. only. We could not identify the CS on two iNPH volumes owing to the severely distorted anatomy, and we excluded their Dice coefficients from subsequent statistical analyses.

Terminology.—We considered BC, CE, SS, and VS to be large classes, and on the primary dataset they represent 10.8%, 2.3%, 1.3%, and 0.6% of all voxels, respectively. The remaining structures were considered small classes. The small classes represented 0.01%–0.28% of all voxels; 85.5% of voxels were background.

Model Architecture

Our U-Net consisted of four downsampling steps, which transformed a two-dimensional 512 × 512 × 1 input image into a 32 × 32 × 1024 feature tensor, and four upsampling steps, which transformed this tensor into a 512 × 512 × 12 segmentation mask (where the third channel corresponds to the 11 ROIs plus background). The model contained 31 042 499 trainable parameters. After experimenting with various configurations, our final model differed from the U-Net developed by Ronneberger through the use of layer normalization (21) between every 3 × 3 convolution and rectified linear unit activation function. We used softmax with categorical cross-entropy loss for multiclass prediction. A schematic of our U-Net can be found in Figure E1 (supplement).

Model Training

Weights from all training sessions were randomly initialized using the method reported by He et al (22). We used the Adam optimizer (23) with an initial learning rate of 0.0001. Data aug-

Table 1: List of Model Configurations Evaluated on the Validation Dataset

Model	Loss Function	Weighting Scheme	Within-Network Normalization
1	Categorical cross-entropy loss	Balanced weighting	No normalization
2	Categorical cross-entropy loss	Balanced weighting	Layer normalization
3*	Categorical cross-entropy loss	Attenuated weighting	Layer normalization
4	Categorical cross-entropy loss	No weighting	Layer normalization
5	Dice loss	No weighting	Layer normalization
6	Categorical cross-entropy loss	Attenuated weighting	Batch normalization (batch size 3)
7	Categorical cross-entropy loss	Attenuated weighting	Batch normalization (batch size 12)
8	Categorical cross-entropy loss	Attenuated weighting	Batch renormalization

* Final model that was used for subsequent analyses.

mentation included left-right flipping, anteroposterior flipping (to accommodate both "left, anterior, superior" and "left, posterior, superior" orientations), scaling (±8%), rotation (±7°), and translation (±20 voxels in x and y). We monitored validation loss and per-class soft Dice coefficients (24), which were the Dice coefficients of predicted softmax class probabilities before thresholding. We stopped training when the average soft Dice coefficient decreased by less than 0.001 over five consecutive epochs. Training was performed on Keras 2.2.4 (Google, Mountain View, Calif) using TensorFlow 1.12 (*https://www.tensorflow.org/*). We used an 11 GB NVIDIA GTX 1080Ti graphics processing unit (Nvidia, Santa Clara, Calif) with a batch size of three.

Experiments

We sequentially evaluated several models on the validation dataset to determine the optimal configuration (Table 1). We started by using the Ronneberger U-Net with Hounsfield units as input (model 1). Subsequently, we added layer normalization between every 3 × 3 convolution and rectified linear unit activation (model 2). With layer normalization in place, we compared three weighting schemes for categorical cross-entropy loss: balanced weighting (model 2) given by

$$w_c = \frac{\sum_{i=0}^{k} n_i}{(k+1)n_c},$$

where w_c is the weight of the cth class, n_c and n_i are the total number of voxels that belong to the cth and ith classes in the entire dataset, respectively (with 0 being background by convention), and k is the total number of structures in the dataset (11 in our case); attenuated weighting given by $w_{attenuated} = \sqrt[3]{w_{balanced}}$ (model 3); and no class weighting (model 4). We implemented cube root because it reduced larger weights to a proportionately greater extent. We also compared Dice loss (model 5) with categorical cross-entropy loss (model 3). We did not use class weights with Dice loss because the Dice coefficient is invariant to the level of spatial representation. Last, we compared layer normalization (21), batch normalization (25) trained on batch sizes of three and 12 (models 6 and 7, respectively), and batch renormalization (model 8) (26). On the basis of validation dataset performance, we selected the most optimal model (model 3) and retrained it on data stratified by scan manufacturer (Appendix E1 [supplement]).

Comparison with Existing Methodology

We compared our model to a publicly available tool (27) using the iNPH dataset. This algorithm registered head CT scans to the MNI152 space and performed segmentation using support vector machine and random forest classifiers. To match output labels, we merged CS with SS, SP with VS, and the remaining structures into "brain parenchyma." Only a portion of each volume was used for inference (12 volumes with 217 slices) because the algorithm was designed to segment images containing the lateral ventricles. Registration failed for one of the 12 volumes because the vendor-specific DICOM coordinates were not properly aligned in the sagittal plane. We replaced this with another study acquired from the same patient using a different scanner at the next nearest date.

Statistical Analysis

We compared the model's structure segmentation Dice coefficients and total volume metrics with the observers using ground truth masks as a reference to determine if the model's segmentation differed from that of domain experts. We compared the model's structure segmentation Dice coefficients on the primary and secondary test datasets to determine if the model's segmentation differed across datasets. Statistical significance was determined using categorical linear regression with $P < .05$ in IBM SPSS statistics version 26 (IBM, Armonk, NY).

Model Availability

The trained model is available at *https://jasonccai.github.io/HeadCTSegmentation/*.

Results

Validation Dataset

Effect of within-network normalization.—The performance of the Ronneberger U-Net (model 1) varied between random initializations, from no learning to learning the large classes (eg, BC). Normalizing intermediate layer activations

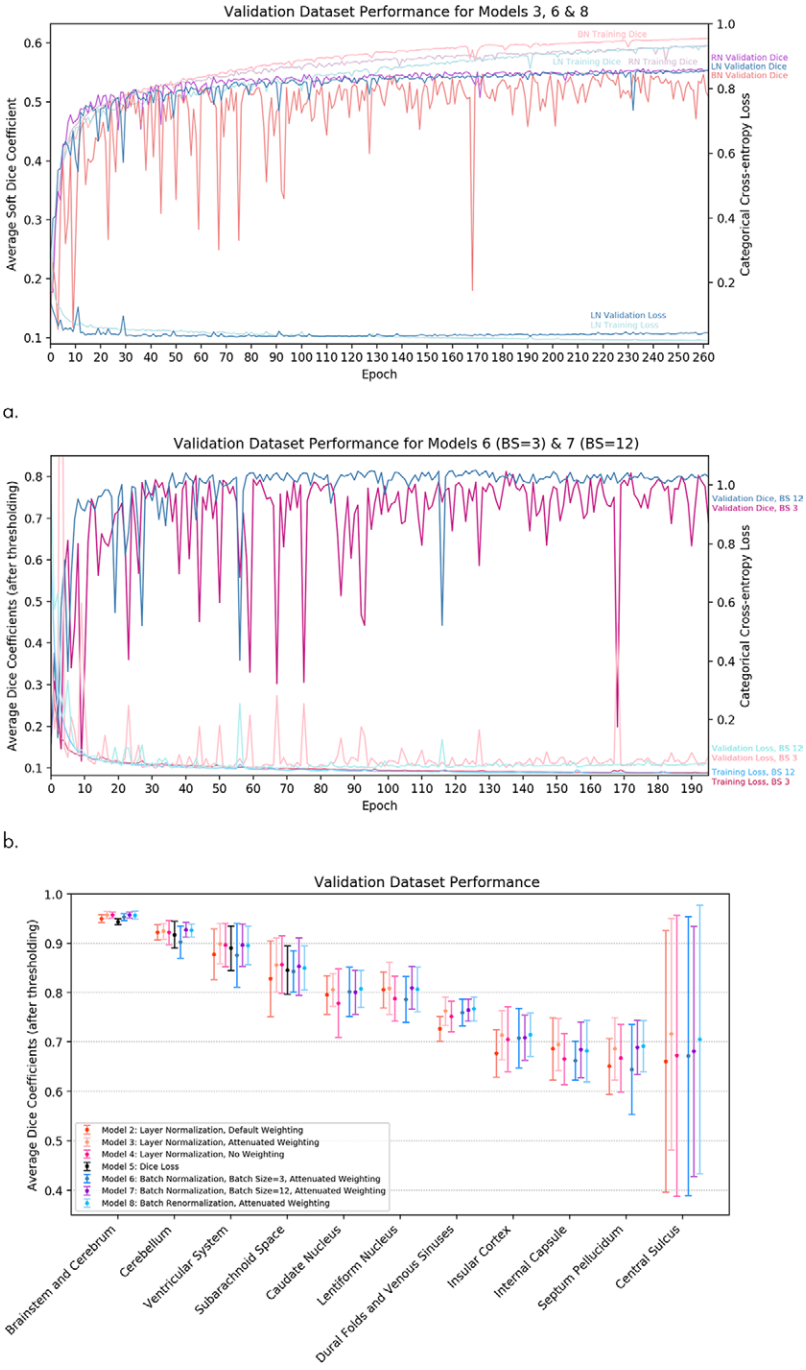

Figure 2: Training and validation performance of various model configurations. **(a)** Batch normalization (BN) demonstrated fluctuations in validation performance. **(b)** This was alleviated in part by increasing the batch size (BS) from three to 12. **(c)** Per-class Dice coefficients averaged over all examinations in the validation dataset. Error bars represent 1 standard deviation. Model 5 predicted only the brainstem and cerebrum, cerebellum, ventricular system, and subarachnoid space. RN = batch renormalization, LN = layer normalization.

enabled the U-Net to predict all 11 structures. With batch normalization, validation performance fluctuated at the start of training (Fig 2). The fluctuation was alleviated in part by increasing the batch size from three to 12 (models 6 and 7, respectively) and circumvented by using batch renormalization (model 8) or layer normalization (model 3). Models that employ layer normalization, batch normalization, and batch renormalization delivered overall Dice coefficients of 0.80 ± 0.06 (standard deviation), 0.79 ± 0.06, and 0.80 ± 0.06, respectively (Table E2 [supplement]).

Effect of loss functions.—Dice loss (model 5) has been reported to perform well on binary segmentation tasks (24). However, when we trained our model using Dice loss, it did not learn smaller classes such as CS, SP, and structures at the basal ganglia level. Instead, they were predicted as SS, VS, and BC, respectively. Figure 3 shows confusion matrices for Dice and cross-entropy loss on the validation dataset.

Effect of class weighting.—When we used balanced class weights in model 2, the model reached an overall Dice coefficient of 0.78 ± 0.06 (Table E2 [supplement]). When we applied an element-wise cube root to these weights in model 3, the overall Dice coefficient improved slightly to 0.80 ± 0.06. Although differences fell within 1 standard deviation, model 3 delivered qualitatively better visual performance (Fig 4, *C*). Model 2 tended to overpredict all 11 ROIs; this manifested additionally as a thin layer of labeled voxels at the brain-cranium boundary (Fig 4, *F*), which was not present on manual segmentations (Fig 4, *E*). When trained without class weighting, model 4 eventually reached an overall Dice coefficient of 0.79 ± 0.06, which was comparable to the performance of model 3. However, model 4 took substantially longer to train. It also learned large classes at the expense of small classes and demonstrated fluctuations in validation performance at the start of training (Fig 4, *A*).

Primary Test Dataset

When structure volume was used to evaluate performance (Table 2), there was no difference between the model and all three observers in BC, CO, DV, LN, SP, and VS; the model was not different from two observers in CA, CE, CN, and SS, and from one observer in CS ($P > .05$). When using Dice coefficients to evaluate performance (Table 3), the model's segmentation was not different from two of the three observers in CE, CO, DV, LN, SP, SS, and VS, and it was not different from one of the three observers in BC and CA ($P > .05$). The model did not perform as well as observers in CN and CS ($P < .05$). The average overall difference in segmentation performance was 4.0% for both measures. Samples of the observers' annotations, ground truth masks, and model predictions are shown in Figure 5.

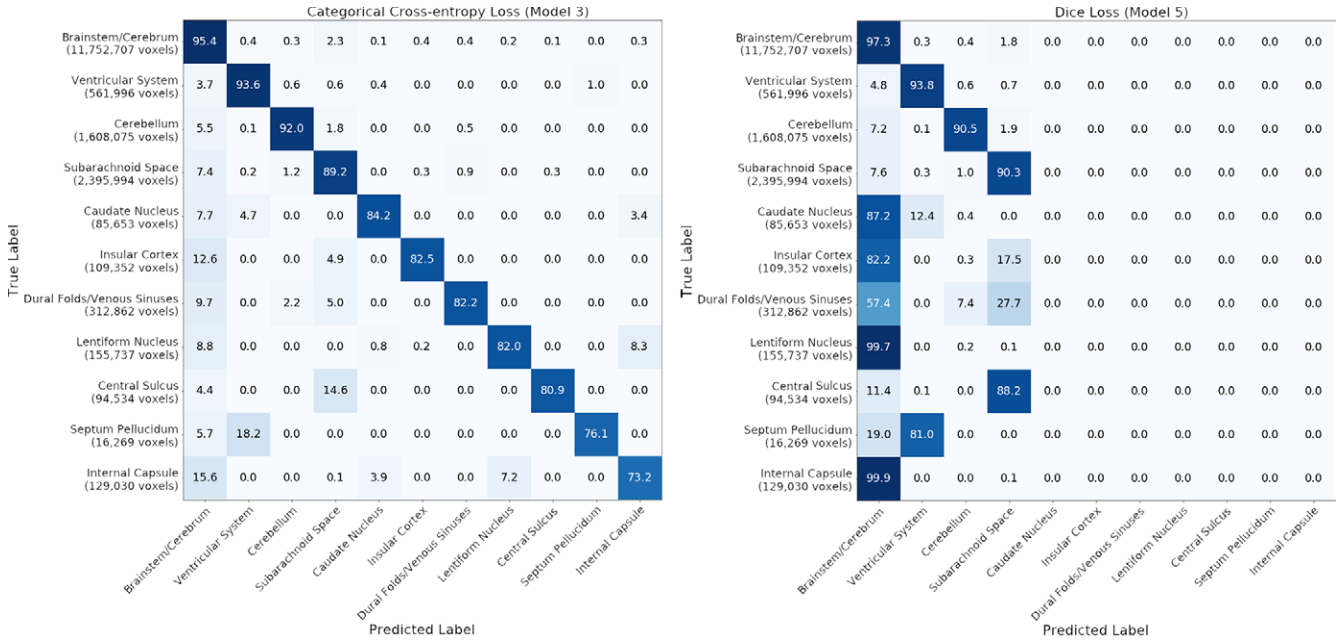

Figure 3: Confusion matrices for **(a)** model 3 (cross-entropy loss with attenuated weighting) and **(b)** model 5 (Dice loss) on the validation dataset. Numbers represent the percentage of voxels in each category.

Secondary Test Datasets

There was no difference in the model's performance between the RSNA dataset and ground truth masks in seven of 11 structures ($P > .05$); the model did not perform as well in BC, SP, SS, and VS ($P < .05$). There was no difference in the model's performance between the iNPH dataset and the primary test dataset in 10 of 11 structures ($P > .05$). The model did not perform as well in segmenting the SP ($P < .05$). Table 4 and Figure 6 show performance on the secondary test dataset.

We observed that the model was slightly better at segmenting the VS of patients with iNPH, although it was not trained on scans demonstrating ventricular enlargement. We also observed a positive correlation between structure size and Dice coefficients in general. This was examined in Appendix E2 (supplement).

Comparison with Existing Methodology

Our model produced the same overall Dice coefficients as a three-dimensional CNN that segmented the SS (9). It demonstrated higher Dice coefficients over a traditional approach (27) ($P < .05$) (Table E3 [supplement]).

Discussion

Several studies have shown that CT structural imaging features are predictive of neurologic disease and patient outcomes (14,15,29–31). However, the inability to rapidly and accurately segment neuroanatomy has impeded the routine clinical use of volumetric imaging features, as well as the discovery of new associations and normative population metrics for these features (16). To address this need, we developed a deep learning model that segmented 11 intracranial structures from noncontrast CT, and we tested the model on internal and external datasets. As these structures ranged considerably in size and slicewise prevalence, we explored the use of within-network normalization techniques, class weighting schemes, and loss functions to obtain the best configuration. These optimizations have general utility for future deep learning–based approaches to medical image segmentation.

The model's performance was not different from observers for most structures, although the small number of samples tested may have reduced our ability to detect significant differences. However, the finding of an overall difference in Dice coefficient of 4% supports the concept that the model exhibited performance similar to that of humans. This difference was larger only for CE and CS, likely as a consequence of the two-dimensional design of the model. In particular, CS is identified by scrolling through the volume. Therefore, the inability of the model to identify features across the z-axis may have hindered its performance. Future work should include examining the possible incremental benefits of three-dimensional and multiplanar approaches to segmentation.

In our experiments, within-network normalization techniques facilitated multiclass learning. Batch normalization has been shown to improve performance by reducing internal covariate shift (25) and stabilizing intermediate activations (32). However, we observed that batch normalization produced significant fluctuations in validation metrics at the start of training. These random fluctuations may be attributed to its dependence on minibatch statistics. Batch normalization estimated the population mean and variance from moving averages calculated during training. With small minibatches (as in our case), such estimates likely became less accurate, resulting in poorer predictions. Increasing the minibatch size from three to 12 reduced but did not resolve this issue, suggesting that even larger minibatches may help. Because of its dependency on large, representative minibatches, batch normalization was less robust on our model.

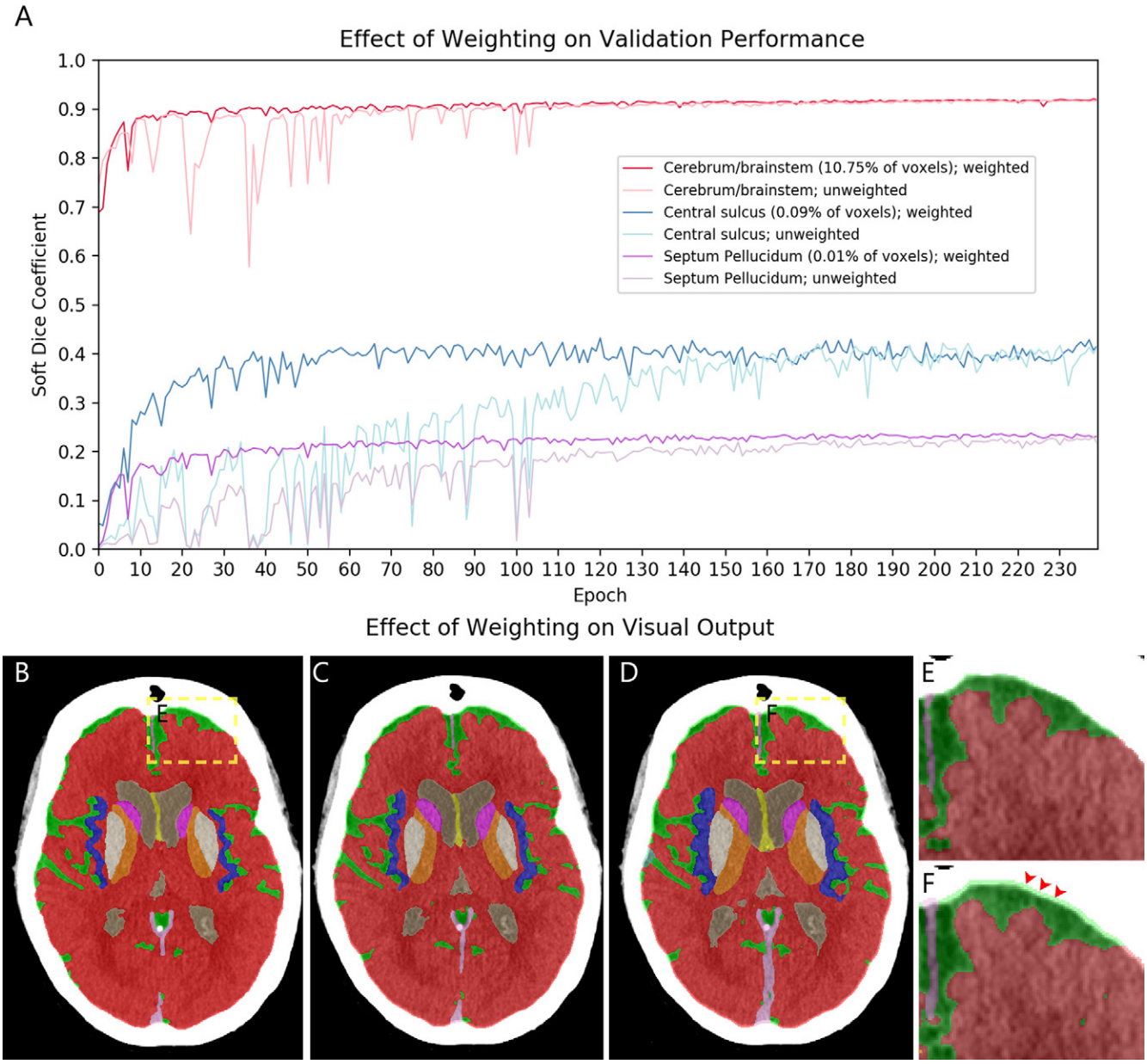

Figure 4: *A,* Without class weights, the model quickly learned large classes (brainstem and cerebrum) at the expense of small classes (central sulcus and septum pellucidum). However, it eventually converged when given sufficient training time. The y-axis represents soft Dice coefficients, which are the Dice coefficients of predicted softmax class probabilities before thresholding. *B,* Manual segmentation, *C,* predictions from model trained with attenuated weighting (model 3), *D,* predictions from model trained with balanced weighting (model 2), *E,* enlarged section from *B* (yellow dotted lines), and, *F,* enlarged section from *D* (yellow dotted lines). The red arrowheads in *F* indicate a thin layer of labeled voxels over the brain-cranium boundary. This thin layer was not seen in the unweighted model. See the section on "Effect of Class Weighting" for details.

Several alternatives have been proposed to address the limitations of batch normalization, including batch renormalization (26) and layer normalization (21). The overall Dice coefficients between these techniques were similar. We preferred layer normalization because it does not depend on minibatch size, does not introduce new hyperparameters, and does not contain parameters that must be learned during training.

Class imbalance degrades CNN performance, as models that learn from imbalanced datasets are biased to reflect the frequency of training samples (33). The soft Dice loss was proposed to reduce model bias for large structures (24). However, as shown on our dataset, the model failed to learn small structures when trained using soft Dice loss, suggesting that it did not entirely recapitulate properties of the Dice coefficient. Instead, we found that weighted categorical cross-entropy loss, which assigned higher misclassification penalties to infrequently represented classes, was better suited to imbalanced multiclass segmentation. This approach balanced gradient updates between classes, shortened overall training time, and promoted a more optimal convergence on our dataset.

A limitation of our work was that we trained our model on a small set of normal head CT scans. Although our model demonstrated good segmentation metrics on iNPH and external scans, it may not generalize equally well to other abnormal scans or

Table 2: Absolute Difference in Each Structure's Volume between the Model's Prediction and Observers' Annotations with that of Ground Truth Masks

Structure	Differences in Volume between Model and GTM (cm^3)	Differences in Volume between Obs1 and GTM (cm^3)	Differences in Volume between Obs2 and GTM (cm^3)	Differences in Volume between Obs3 and GTM (cm^3)	Average Total Volume in GTM (cm^3)	$\lvert\frac{\sum \text{observers}1,2,3}{3} - \text{model}$ (% of total volume)
Brainstem and cerebrum	4.89 ± 3.71	4.44 ± 4.16	3.89 ± 2.91	4.03 ± 2.26	177.99 ± 25.37	0.77 (0.44)
Caudate nucleus	0.27 ± 0.19	0.14 ± 0.12*	0.16 ± 0.14	0.22 ± 0.12	3.00 ± 0.47	0.10 (3.33)
Central sulcus	0.29 ± 0.32	0.08 ± 0.09*	0.14 ± 0.13	0.08 ± 0.07*	1.13 ± 0.57	0.19 (16.81)
Cerebellum	3.43 ± 2.52	2.55 ± 3.50	1.24 ± 1.49*	2.30 ± 2.67	16.36 ± 9.71	1.40 (8.56)
Dural folds and venous sinuses	0.65 ± 0.52	0.46 ± 0.36	1.08 ± 1.24	0.43 ± 0.46	3.79 ± 1.35	0.01 (0.26)
Insular cortex	0.45 ± 0.50	0.46 ± 0.43	0.37 ± 0.30	0.31 ± 0.35	4.50 ± 0.81	0.06 (1.33)
Internal capsule	0.73 ± 0.44	0.63 ± 0.19	0.42 ± 0.18*	0.53 ± 0.39	4.64 ± 0.82	0.21 (4.53)
Lentiform nucleus	0.57 ± 0.86	0.45 ± 0.45	0.50 ± 0.47	0.53 ± 0.40	6.78 ± 1.23	0.08 (1.18)
Septum pellucidum	0.10 ± 0.09	0.14 ± 0.14	0.17 ± 0.12	0.06 ± 0.09	0.48 ± 0.22	0.03 (6.25)
Subarachnoid space	3.61 ± 1.87	1.61 ± 0.97*	3.83 ± 2.77	4.97 ± 2.46	29.20 ± 10.67	0.15 (0.51)
Ventricular system	0.46 ± 0.30	0.58 ± 0.88	0.37 ± 0.29	0.47 ± 0.43	14.17 ± 9.30	0.02 (0.14)
Overall	1.40 ± 1.03	1.05 ± 1.03	1.11 ± 0.91	1.26 ± 0.88	NA	0.26 (3.94)

Note.—Values shown are mean ± standard deviation or mean with percentages in parentheses. In the farthest column to the right (% of total volume), figures do not represent anatomic volumes because only four to six slices were sampled from each examination to generate ground truth masks (GTMs). The comparison methodology is outlined in Figure 1. NA = not applicable, Obs = observer.
* Indicates higher performance compared with the model ($P < .05$).

Table 3: Average Dice Coefficients of Each Structure with Reference to the Ground Truth Masks

Structure	Model vs GTM	Obs1 vs GTM	Obs2 vs GTM	Obs3 vs GTM	$\lvert\frac{\sum \text{observers}1,2,3}{3} - \text{model}$
Brainstem and cerebrum	0.95 ± 0.01	0.95 ± 0.01	0.96 ± 0.01*	0.96 ± 0.01*	0.01
Caudate nucleus	0.89 ± 0.03	0.92 ± 0.02*	0.94 ± 0.02*	0.93 ± 0.03*	0.04
Central sulcus	0.79 ± 0.13	0.88 ± 0.06*	0.86 ± 0.05*	0.93 ± 0.03*	0.10
Cerebellum	0.83 ± 0.07	0.89 ± 0.13	0.93 ± 0.05*	0.90 ± 0.11	0.07
Dural folds and venous sinuses	0.76 ± 0.09	0.81 ± 0.07	0.78 ± 0.09	0.85 ± 0.07*	0.05
Insular cortex	0.81 ± 0.09	0.81 ± 0.05	0.85 ± 0.03	0.87 ± 0.06*	0.04
Internal capsule	0.77 ± 0.06	0.76 ± 0.07	0.86 ± 0.04*	0.82 ± 0.06*	0.04
Lentiform nucleus	0.87 ± 0.06	0.86 ± 0.05	0.93 ± 0.04*	0.91 ± 0.04	0.03
Septum pellucidum	0.75 ± 0.06	0.66 ± 0.25	0.76 ± 0.14	0.89 ± 0.08*	0.02
Subarachnoid space	0.84 ± 0.03	0.84 ± 0.03	0.85 ± 0.01	0.89 ± 0.03*	0.02
Ventricular system	0.91 ± 0.03	0.92 ± 0.03	0.92 ± 0.03	0.95 ± 0.02*	0.02
Overall	0.83 ± 0.06	0.85 ± 0.07	0.88 ± 0.05	0.90 ± 0.05	0.04

Note.—Values shown are mean ± standard deviation. A graphical representation of this table is shown in Figure 5. The comparison methodology is outlined in Figure 1. GTM = ground truth mask, Obs = observer.
* Indicates higher performance compared with the model ($P < .05$).

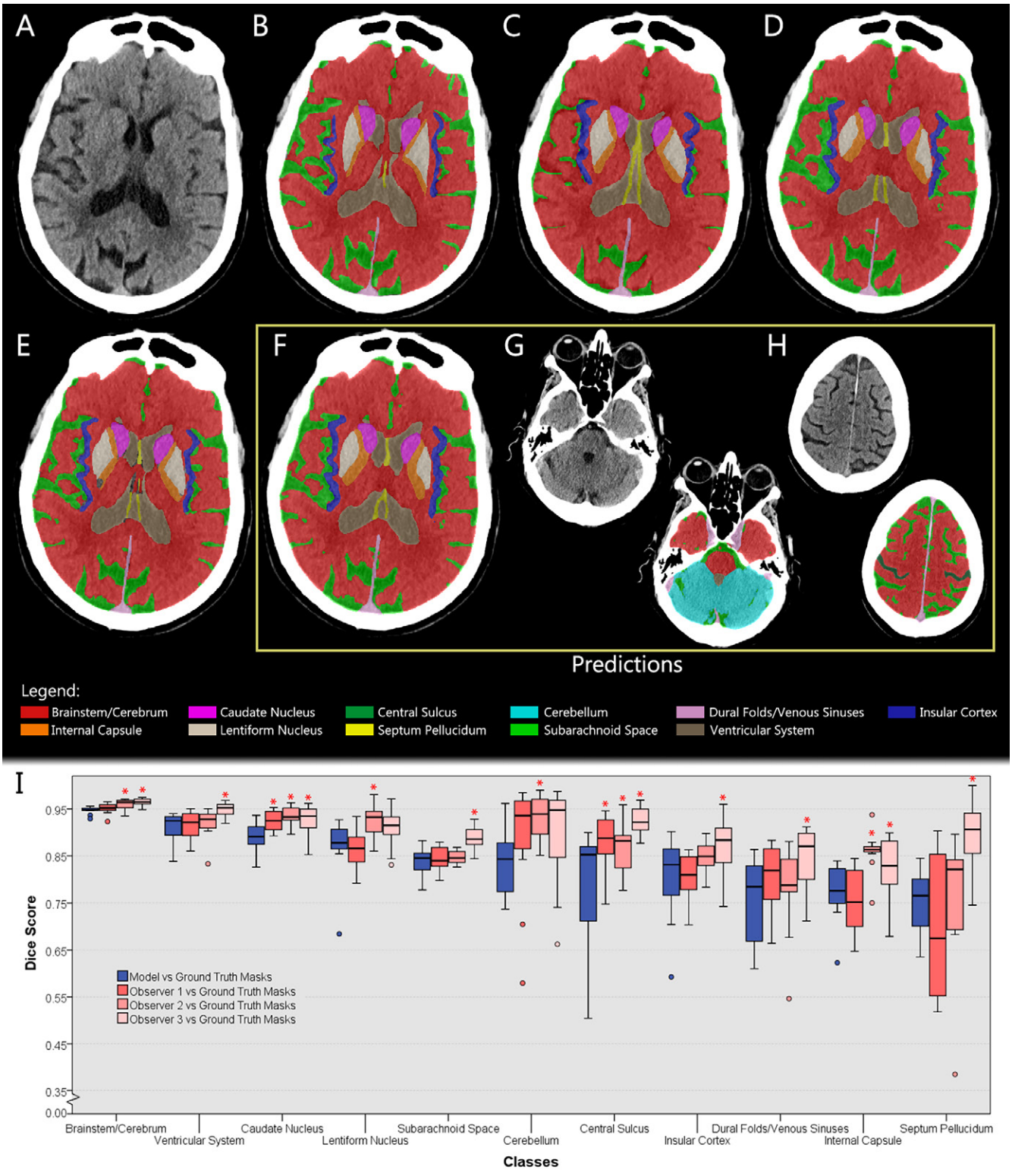

Figure 5: Sample image from the primary test dataset with corresponding model prediction, expert annotations, and ground truth mask. Images shown are, *A*, original image, *B*, observer 1 segmentation, *C*, observer 2 segmentation, *D*, observer 3 segmentation, *E*, ground truth mask generated by majority voting (unlabeled voxels indicate areas where all three observers disagree and are excluded from performance metrics), *F*, model prediction at basal ganglia level, *G*, model prediction at cerebellar level, and, *H*, model prediction at central sulcus level. *I*, Box and whisker plot of Dice coefficients using ground truth masks as reference ($n = 59$ slices). Asterisks indicate statistically significant results as compared with the model ($P < .05$). Dice coefficients are also presented in Table 3. The comparison methodology is outlined in Figure 1.

Table 4: Performance between Primary and Secondary Test Datasets

	Comparison 1: Both Datasets Contain Fully Annotated Volumes		Comparison 2: Both Datasets Contain Four to Six Representative Slices from each Volume	
Structure	iNPH Dataset	Primary Test Dataset	RSNA Dataset	Ground Truth Masks
Brainstem and cerebrum	0.96 ± 0.01	0.96 ± 0.01	0.93 ± 0.02	0.95 ± 0.01*
Caudate nucleus	0.83 ± 0.06	0.84 ± 0.05	0.88 ± 0.03	0.89 ± 0.03
Central sulcus	0.50 ± 0.35	0.77 ± 0.12	0.74 ± 0.23	0.79 ± 0.13
Cerebellum	0.92 ± 0.04	0.94 ± 0.02	0.86 ± 0.10	0.83 ± 0.07
Dural folds and venous sinuses	0.78 ± 0.03	0.80 ± 0.04	0.70 ± 0.08	0.76 ± 0.09
Insular cortex	0.81 ± 0.07	0.77 ± 0.08	0.75 ± 0.12	0.81 ± 0.09
Internal capsule	0.76 ± 0.05	0.74 ± 0.06	0.77 ± 0.08	0.77 ± 0.06
Lentiform nucleus	0.87 ± 0.04	0.85 ± 0.05	0.86 ± 0.06	0.87 ± 0.06
Septum pellucidum	0.68 ± 0.08	0.76 ± 0.08*	0.63 ± 0.11	0.75 ± 0.06*
Subarachnoid space	0.86 ± 0.03	0.87 ± 0.04	0.77 ± 0.06	0.84 ± 0.03*
Ventricular system	0.95 ± 0.01*	0.91 ± 0.05	0.87 ± 0.04	0.91 ± 0.03*
Overall	0.81 ± 0.07	0.84 ± 0.05	0.80 ± 0.09	0.83 ± 0.06

Note.—Values shown are mean ± standard deviation of Dice coefficients. A graphical representation of this table is shown in Figure 6. The comparison methodology is outlined in Figure 1. iNPH = idiopathic normal pressure hydrocephalus, RSNA = Radiological Society of North America.
* Indicates higher performance ($P < .05$).

imaging acquired using different protocols because the training dataset may not have fully captured the diversity of head CT scans. Transfer learning with fine-tuning is a well-established technique that enables deep learning models to be quickly optimized on new and previously unseen datasets. In making our model available to the artificial intelligence community, we hope that developers can use transfer learning to optimize it for their specific needs.

In conclusion, given the increasing amount of imaging data, algorithms that automatically extract quantitative information and spatial context from head CT can potentially assist in clinical decision making. In this study, we trained and validated a U-Net to simultaneously segment 11 intracranial structures on head CT. Our model correlated well with expert annotations, and it delivered equal or higher overall Dice coefficients as compared with existing automated segmentation methods. We showed that within-network normalization facilitated multiclass learning, and the choice of normalization technique affected the network's overall robustness and stability. We also demonstrated the effects of loss functions and class weighting schemes on network performance. Our findings have general utility for future deep learning–based approaches to medical image segmentation. By adapting the U-Net for multiclass segmentation on head CT scans, we also set a framework for similar work in the future and further affirm the role of deep learning in supporting radiologists.

Acknowledgments: We would like to thank Mayo Clinic for funding this study. The RSNA Intracranial Hemorrhage Detection Challenge is a publicly available dataset made possible by the Radiological Society of North America.

Author contributions: Guarantors of integrity of entire study, J.C.C., Z.A., A.B., A.Z., B.J.E.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, J.C.C. Z.A., K.A.P., A.B., A.Z., D.C.V., B.J.E.; clinical studies, Z.A., S.H., A.Z.; experimental studies, J.C.C., Z.A., K.A.P., A.D.W., D.C.V.; statistical analysis, J.C.C., Z.A., K.A.P., A.B., P.R., G.M.C., A.Z., B.E.; and manuscript editing, J.C.C., Z.A., K.A.P., A.B., S.H., A.D.W., P.R., G.M.C., A.Z., Q.H., B.J.E.

Disclosures of Conflicts of Interest: **J.C.C.** disclosed no relevant relationships. **Z.A.** disclosed no relevant relationships. **K.A.P.** disclosed no relevant relationships. **A.B.** disclosed no relevant relationships. **S.H.** disclosed no relevant relationships. **A.D.W.** disclosed no relevant relationships. **P.R.** disclosed no relevant relationships. **G.M.C.** disclosed no relevant relationships. **A.Z.** disclosed no relevant relationships. **D.C.V.** disclosed no relevant relationships. **Q.H.** disclosed no relevant relationships. **B.J.E.** disclosed no relevant relationships.

References

1. Rosman DA, Duszak R Jr, Wang W, Hughes DR, Rosenkrantz AB. Changing Utilization of Noninvasive Diagnostic Imaging Over 2 Decades: An Examination Family-Focused Analysis of Medicare Claims Using the Neiman Imaging Types of Service Categorization System. AJR Am J Roentgenol 2018;210(2):364–368.
2. Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD. Deep Learning in Radiology: Does One Size Fit All? J Am Coll Radiol 2018;15(3 Pt B):521–526.
3. Akkus Z, Cai J, Boonrod A, et al. A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence-Powered Ultrasound for Improving Clinical Workflow. J Am Coll Radiol 2019;16(9 Pt B):1318–1328.
4. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. J Digit Imaging 2017;30(4):449–459.
5. Islam M, Sanghani P, See AAQ, James ML, King NKK, Ren H. ICHNet: Intracerebral Hemorrhage (ICH) Segmentation Using Deep Learning. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, eds. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham, Switzerland: Springer, 2019; 456–463.
6. Chang PD, Kuoy E, Grinband J, et al. Hybrid 3D/2D convolutional neural network for hemorrhage evaluation on head CT. AJNR Am J Neuroradiol 2018;39(9):1609–1616.

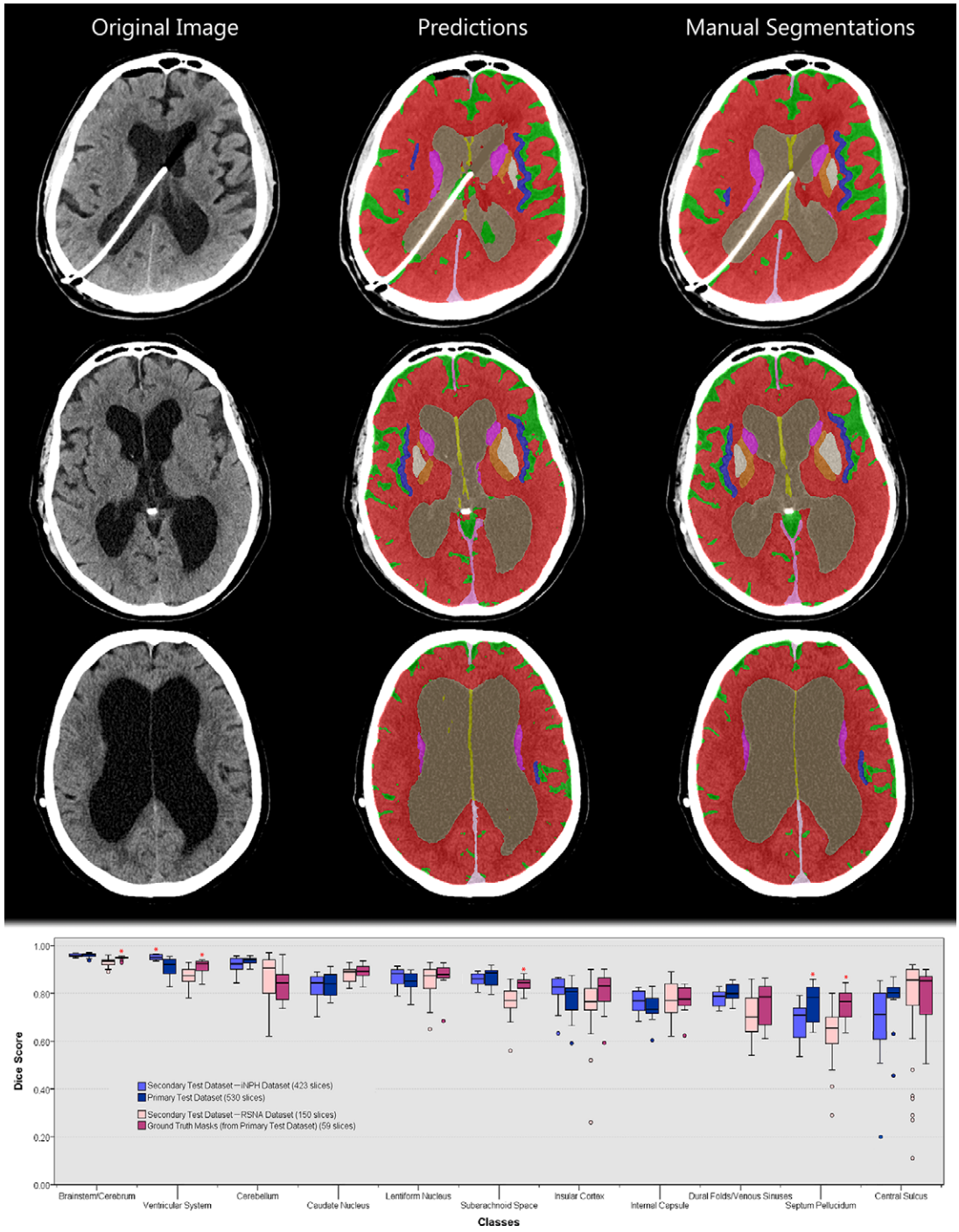

Figure 6: Top: Sample images from the idiopathic normal pressure hydrocephalus (iNPH) dataset. The model was not trained on iNPH scans. Coloring legend can be found in Figure 5. Bottom: Box and whisker plot comparing model performance between primary and secondary test datasets. Asterisks indicate statistically significant results ($P < .05$). Dice coefficients are also presented in Table 4. The comparison methodology is outlined in Figure 1.

7. Gao XW, Qian Y. Segmentation of brain lesions from CT images based on deep learning techniques. In: Gimi B, Krol A, eds. Proceedings of SPIE: medical imaging 2018—biomedical applications in molecular, structural, and functional imaging. Vol 10578. Bellingham, Wash: International Society for Optics and Photonics, 2018; 105782L.
8. Akkus Z, Kostandy P, Philbrick KA, Erickson BJ. Robust brain extraction tool for CT head images. Neurocomputing 2019;392:189–195.
9. Patel A, van de Leemput SC, Prokop M, van Ginneken B, Manniesing R. Automatic cerebrospinal fluid segmentation in non-contrast CT images using a 3D convolutional network. In: Armato SG III, Petrick NA, eds. Proceedings of SPIE: medical imaging 2017—computer-aided diagnosis. Vol 10134. Bellingham, Wash: International Society for Optics and Photonics, 2017; 1013420.
10. Despotović I, Goossens B, Philips W. MRI segmentation of the human brain: challenges, methods, and applications. Comput Math Methods Med 2015;2015:450341.
11. Takahashi N, Shinohara Y, Kinoshita T, et al. Computerized identification of early ischemic changes in acute stroke in noncontrast CT using deep learning. In: Mori K, Hahn HK, eds. Proceedings of SPIE: medical imaging 2019—computer-aided diagnosis. Vol 10950. Bellingham, Wash: International Society for Optics and Photonics, 2019; 109503A.

12. Fritscher KD, Peroni M, Zaffino P, Spadea MF, Schubert R, Sharp G. Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours. Med Phys 2014;41(5):051910.
13. Golby AJ, ed. Image-Guided Neurosurgery. Amsterdam, the Netherlands: Elsevier Science, 2015.
14. Diprose WK, Diprose JP, Wang MTM, Tarr GP, McFetridge A, Barber PA. Automated Measurement of Cerebral Atrophy and Outcome in Endovascular Thrombectomy. Stroke 2019;50(12):3636–3638.
15. Anderson RC, Grant JJ, de la Paz R, Frucht S, Goodman RR. Volumetric measurements in the detection of reduced ventricular volume in patients with normal-pressure hydrocephalus whose clinical condition improved after ventriculoperitoneal shunt placement. J Neurosurg 2002;97(1):73–79.
16. Toma AK, Holl E, Kitchen ND, Watkins LD. Evans' index revisited: the need for an alternative in normal pressure hydrocephalus. Neurosurgery 2011;68(4):939–944.
17. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv e-prints [preprint]. https://arxiv.org/abs/1505.04597. Posted 2015. Accessed October 2019.
18. Paciaroni M, Silvestrelli G, Caso V, et al. Neurovascular territory involved in different etiological subtypes of ischemic stroke in the Perugia Stroke Registry. Eur J Neurol 2003;10(4):361–365.
19. Barber PA, Demchuk AM, Zhang J, Buchan AM. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. ASPECTS Study Group. Alberta Stroke Programme Early CT Score. Lancet 2000;355(9216):1670–1674.
20. Philbrick KA, Weston AD, Akkus Z, et al. RIL-Contour: a Medical Imaging Dataset Annotation Tool for and with Deep Learning. J Digit Imaging 2019;32(4):571–581.
21. Lei Ba J, Kiros JR, Hinton GE. Layer Normalization. arXiv e-prints [preprint]. https://arxiv.org/abs/1607.06450. Posted 2016. Accessed October 2019.
22. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv e-prints [preprint]. https://arxiv.org/abs/1502.01852. Posted 2015. Accessed October 2019.
23. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv e-prints [preprint]. https://arxiv.org/abs/1412.6980. Posted 2014. Accessed October 2019.
24. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. arXiv e-prints [preprint]. https://arxiv.org/abs/1606.04797. Posted 2016. Accessed October 2019.
25. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv e-prints [preprint]. https://arxiv.org/abs/1502.03167. Posted 2015. Accessed October 2019.
26. Ioffe S. Batch Renormalization: Towards Reducing Minibatch Dependence in Batch-Normalized Models. arXiv e-prints [preprint]. https://arxiv.org/abs/1702.03275. Posted 2017. Accessed October 2019.
27. Zhang A, Kao PY, Sahyouni R, Shelat A, Chen J, Manjunath BS. Automated Segmentation of CT Scans for Normal Pressure Hydrocephalus. arXiv e-prints [preprint]. arXiv:1901.09088. https://arxiv.org/abs/1901.09088. Posted 2019. Accessed October 2019.
28. Irimia A, Maher AS, Rostowsky KA, Chowdhury NF, Hwang DH, Law EM. Brain Segmentation From Computed Tomography of Healthy Aging and Geriatric Concussion at Variable Spatial Resolutions. Front Neuroinform 2019;13:9.
29. Frisoni GB, Geroldi C, Beltramello A, et al. Radial width of the temporal horn: a sensitive measure in Alzheimer disease. AJNR Am J Neuroradiol 2002;23(1):35–47.
30. Relkin N, Marmarou A, Klinge P, Bergsneider M, Black PM. Diagnosing idiopathic normal-pressure hydrocephalus. Neurosurgery 2005;57(3 Suppl):S4–S16; discussion ii–v.
31. Kauw F, Velthuis BK, Dankbaar JW. Response by Kauw et al to Letter Regarding Article, "Intracranial Cerebrospinal Fluid Volume as a Predictor of Malignant Middle Cerebral Artery Infarction". Stroke 2019;50(10):e304.
32. Santurkar S, Tsipras D, Ilyas A, Madry A. How Does Batch Normalization Help Optimization? arXiv e-prints [preprint]. https://arxiv.org/abs/1805.11604. Posted 2018. Accessed October 2019.
33. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. arXiv e-prints [preprint]. https://arxiv.org/abs/1710.05381. Posted 2017. Accessed October 2019.